Probability Review

Gonzalo Mateos
Dept. of ECE and Goergen Institute for Data Science
University of Rochester
gmateosb@ece.rochester.edu
http://www.ece.rochester.edu/~gmateosb/

September 20, 2017

Markov and Chebyshev's inequalities

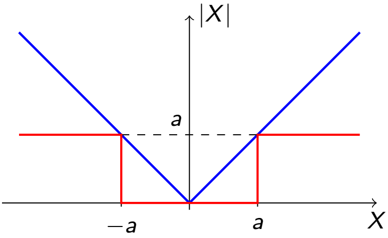
Markov's inequality

- RV X with $\mathbb{E}[|X|] < \infty$, constant $a > 0$
- Markov's inequality states $\Rightarrow \mathrm{P}(|X| \geq a) \leq \dfrac{\mathbb{E}(|X|)}{a}$

Proof.

- $\mathbb{I}\{|X| \geq a\} = 1$ when $|X| \geq a$ and 0 else. Then (figure to the right)

$$a\mathbb{I}\{|X| \geq a\} \leq |X|$$



- Use linearity of expected value

$$a\mathbb{E}(\mathbb{I}\{|X| \geq a\}) \leq \mathbb{E}(|X|)$$

- Indicator function's expectation = Probability of indicated event

$$a\mathrm{P}(|X| \geq a) \leq \mathbb{E}(|X|)$$

□

Chebyshev's inequality

- RV X with $\mathbb{E}(X) = \mu$ and $\mathbb{E}\left[(X-\mu)^2\right] = \sigma^2$, constant $k > 0$
- Chebyshev's inequality states $\Rightarrow \mathrm{P}\left(|X-\mu| \geq k\right) \leq \dfrac{\sigma^2}{k^2}$

Proof.

- Markov's inequality for the RV $Z = (X-\mu)^2$ and constant $a = k^2$

$$\mathrm{P}\left((X-\mu)^2 \geq k^2\right) = \mathrm{P}\left(|Z| \geq k^2\right) \leq \frac{\mathbb{E}\left[|Z|\right]}{k^2} = \frac{\mathbb{E}\left[(X-\mu)^2\right]}{k^2}$$

- Notice that $(X-\mu)^2 \geq k^2$ if and only if $|X-\mu| \geq k$ thus

$$\mathrm{P}\left(|X-\mu| \geq k\right) \leq \frac{\mathbb{E}\left[(X-\mu)^2\right]}{k^2}$$

- Chebyshev's inequality follows from definition of variance

$\square$

Comments and observations

- If absolute expected value is finite, i.e., $\mathbb{E}\left[|X|\right] < \infty$
  - $\Rightarrow$ Complementary (c)cdf decreases at least like x^{-1} (Markov's)
- If mean $\mathbb{E}(X)$ and variance $\mathbb{E}\left[(X-\mu)^2\right]$ are finite
  - $\Rightarrow$ Ccdf decreases at least like x^{-2} (Chebyshev's)
- Most cdfs decrease exponentially (e.g. e^{-x^2} for normal)
  - $\Rightarrow$ Power law bounds $\propto x^{-\alpha}$ are loose but still useful
- Markov's inequality often derived for nonnegative RV $X \geq 0$
  - $\Rightarrow$ Can drop the absolute value to obtain $\mathrm{P}\left(X \geq a\right) \leq \frac{\mathbb{E}(X)}{a}$
  - $\Rightarrow$ General bound $\mathrm{P}\left(X \geq a\right) \leq \frac{\mathbb{E}(X^r)}{a^r}$ holds for $r > 0$

Convergence of random variables

Limits

- Sequence of RVs $X_{\mathbb{N}} = X_1, X_2, \ldots, X_n, \ldots$

  $\Rightarrow$ Distinguish between random process $X_{\mathbb{N}}$ and realizations $x_{\mathbb{N}}$

Q1) Say something about X_n for n large? $\Rightarrow$ Not clear, X_n is a RV

Q2) Say something about x_n for n large? $\Rightarrow$ Certainly, look at $\lim_{n\to\infty} x_n$

Q3) Say something about $\mathrm{P}(X_n \in \mathcal{X})$ for n large? $\Rightarrow$ Yes, $\lim_{n\to\infty} \mathrm{P}(X_n \in \mathcal{X})$

- Translate what we now about regular limits to definitions for RVs
- Can start from convergence of sequences: $\lim_{n\to\infty} x_n$

  $\Rightarrow$ Sure and almost sure convergence

- Or from convergence of probabilities: $\lim_{n\to\infty} \mathrm{P}(X_n)$

  $\Rightarrow$ Convergence in probability, in mean square and distribution

Convergence of sequences and sure convergence

- Denote sequence of numbers $x_{\mathbb{N}} = x_1, x_2, \ldots, x_n, \ldots$
- **Def:** Sequence $x_{\mathbb{N}}$ converges to the value x if given any $\epsilon > 0$
  - $\Rightarrow$ There exists n_0 such that for all $n > n_0$, $|x_n - x| < \epsilon$
- Sequence x_n comes arbitrarily close to its limit $\Rightarrow |x_n - x| < \epsilon$
  - $\Rightarrow$ And stays close to its limit for all $n > n_0$
- Random process (sequence of RVs) $X_{\mathbb{N}} = X_1, X_2, \ldots, X_n, \ldots$
  - $\Rightarrow$ Realizations of $X_{\mathbb{N}}$ are sequences $x_{\mathbb{N}}$
- **Def:** We say $X_{\mathbb{N}}$ converges surely to RV X if
  - $\Rightarrow \lim_{n \to \infty} x_n = x$ for all realizations $x_{\mathbb{N}}$ of $X_{\mathbb{N}}$
- Said differently, $\lim_{n \to \infty} X_n(s) = X(s)$ for all $s \in S$
- Not really adequate. Even a (practically unimportant) outcome that happens with vanishingly small probability prevents sure convergence

Almost sure convergence

- RV X and random process $X_{\mathbb{N}} = X_1, X_2, \ldots, X_n, \ldots$
- **Def:** We say $X_{\mathbb{N}}$ converges almost surely to RV X if

$$\mathrm{P}\left(\lim_{n\to\infty} X_n = X\right) = 1$$

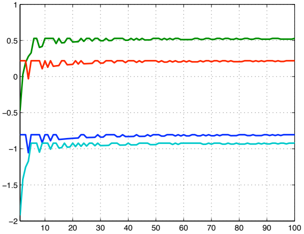
$\Rightarrow$ Almost all sequences converge, except for a set of measure 0

- Almost sure convergence denoted as $\Rightarrow \lim_{n\to\infty} X_n = X \quad \text{a.s.}$

$\Rightarrow$ Limit X is a random variable

Example

- $X_0 \sim \mathcal{N}(0,1)$ (normal, mean 0, variance 1)
- Z_n sequence of Bernoulli RVs, parameter p
- Define $\Rightarrow X_n = X_0 - \dfrac{Z_n}{n}$
- $\dfrac{Z_n}{n} \to 0$ so $\lim_{n\to\infty} X_n = X_0$ a.s. (also surely)



Almost sure convergence example

- Consider $S = [0, 1]$ and let $\mathrm{P}(\cdot)$ be the uniform probability distribution
  $\Rightarrow \mathrm{P}([a, b]) = b - a$ for $0 \leq a \leq b \leq 1$
- Define the RVs $X_n(s) = s + s^n$ and $X(s) = s$
- For all $s \in [0, 1)$ $\Rightarrow s^n \to 0$ as $n \to \infty$, hence $X_n(s) \to s = X(s)$
- For $s = 1$ $\Rightarrow X_n(1) = 2$ for all n, while $X(1) = 1$
- Convergence only occurs on the set $[0, 1)$, and $\mathrm{P}([0, 1)) = 1$
  $\Rightarrow$ We say $\lim_{n\to\infty} X_n = X$ a.s.
  $\Rightarrow$ Once more, note the limit X is a random variable

Convergence in probability

- **Def:** We say $X_{\mathbb{N}}$ converges in probability to RV X if for any $\epsilon > 0$

$$\lim_{n\to\infty} \mathrm{P}\left(|X_n - X| < \epsilon\right) = 1$$

$\Rightarrow$ Prob. of distance $|X_n - X|$ becoming smaller than ϵ tends to 1

- Statement is about probabilities, not about realizations (sequences)
  $\Rightarrow$ Probability converges, realizations $x_{\mathbb{N}}$ may or may not converge
  $\Rightarrow$ Limit and prob. interchanged with respect to a.s. convergence

Theorem

Almost sure (a.s.) convergence implies convergence in probability

Proof.

- If $\lim_{n\to\infty} X_n = X$ then for any $\epsilon > 0$ there is n_0 such that

$$|X_n - X| < \epsilon \text{ for all } n \geq n_0$$

- True for all almost all sequences so $\mathrm{P}\left(|X_n - X| < \epsilon\right) \to 1$

$\square$

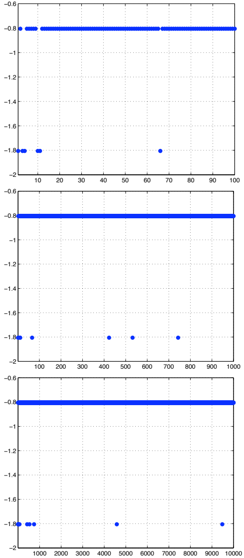
Convergence in probability example

- $X_0 \sim \mathcal{N}(0, 1)$ (normal, mean 0, variance 1)
- Z_n sequence of Bernoulli RVs, parameter $1/n$
- Define $\Rightarrow X_n = X_0 - Z_n$
- X_n converges in probability to X_0 because

$$\begin{aligned}
\mathrm{P}\left(|X_n - X_0| < \epsilon\right) &= \mathrm{P}\left(|Z_n| < \epsilon\right) \\
&= 1 - \mathrm{P}\left(Z_n = 1\right) \\
&= 1 - \frac{1}{n} \to 1
\end{aligned}$$

- Plot of path x_n up to $n = 10^2$, $n = 10^3$, $n = 10^4$

  $\Rightarrow Z_n = 1$ becomes ever rarer but still happens



Difference between a.s. and in probability

- Almost sure convergence implies that almost all sequences converge
- Convergence in probability does not imply convergence of sequences

- Latter example: $X_n = X_0 - Z_n$, Z_n is Bernoulli with parameter $1/n$

  $\Rightarrow$ Showed it converges in probability

$$\mathrm{P}\left(|X_n - X_0| < \epsilon\right) = 1 - \frac{1}{n} \to 1$$

  $\Rightarrow$ But for almost all sequences, $\lim\limits_{n\to\infty} x_n$ does not exist

- Almost sure convergence $\Rightarrow$ disturbances stop happening
- Convergence in prob. $\Rightarrow$ disturbances happen with vanishing freq.
- Difference not irrelevant
  - Interpret Z_n as rate of change in savings
  - With a.s. convergence risk is eliminated
  - With convergence in prob. risk decreases but does not disappear

Mean-square convergence

- **Def:** We say $X_{\mathbb{N}}$ converges in mean square to RV X if

$$\lim_{n\to\infty} \mathbb{E}\left[|X_n - X|^2\right] = 0$$

$\Rightarrow$ Sometimes (very) easy to check

Theorem
Convergence in mean square implies convergence in probability

Proof.

- From Markov's inequality

$$\mathrm{P}\left(|X_n - X| \geq \epsilon\right) = \mathrm{P}\left(|X_n - X|^2 \geq \epsilon^2\right) \leq \frac{\mathbb{E}\left[|X_n - X|^2\right]}{\epsilon^2}$$

- If $X_n \to X$ in mean-square sense, $\mathbb{E}\left[|X_n - X|^2\right]/\epsilon^2 \to 0$ for all ϵ

$\square$

- Almost sure and mean square $\Rightarrow$ neither one implies the other

Convergence in distribution

- Consider a random process $X_{\mathbb{N}}$. Cdf of X_n is $F_n(x)$
- **Def:** We say $X_{\mathbb{N}}$ converges in distribution to RV X with cdf $F_X(x)$ if

  $\Rightarrow \lim_{n\to\infty} F_n(x) = F_X(x)$ for all x at which $F_X(x)$ is continuous

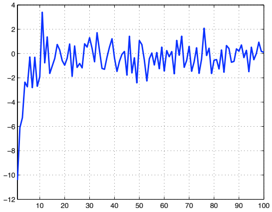
- No claim about individual sequences, just the cdf of X_n

  $\Rightarrow$ Weakest form of convergence covered

- Implied by almost sure, in probability, and mean square convergence

Example

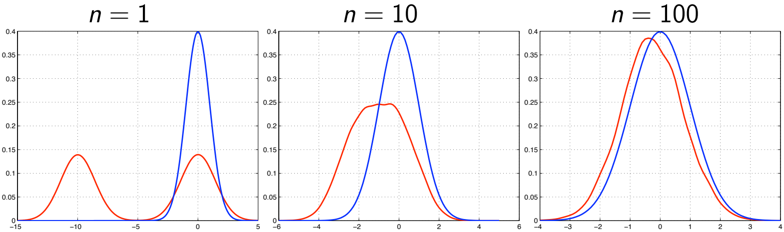
- $Y_n \sim \mathcal{N}(0,1)$
- Z_n Bernoulli with parameter p
- Define $\Rightarrow X_n = Y_n - 10Z_n/n$
- $\dfrac{Z_n}{n} \to 0$ so $\lim_{n\to\infty} F_n(x)$ "$=$" $\mathcal{N}(0,1)$



Convergence in distribution (continued)

- Individual sequences x_n do not converge in any sense

  $\Rightarrow$ It is the distribution that converges

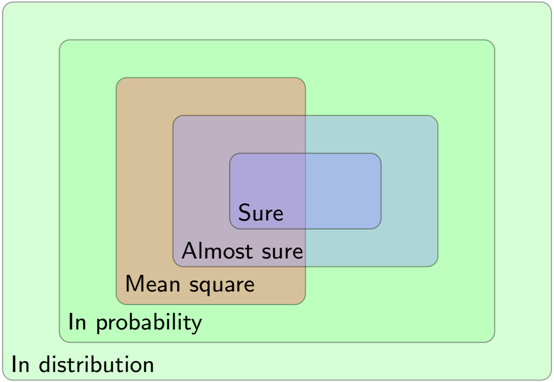


- As the effect of Z_n/n vanishes pdf of X_n converges to pdf of Y_n

  $\Rightarrow$ Standard normal $\mathcal{N}(0, 1)$

Implications

- Sure $\Rightarrow$ almost sure $\Rightarrow$ in probability $\Rightarrow$ in distribution
- Mean square $\Rightarrow$ in probability $\Rightarrow$ in distribution
- In probability $\Rightarrow$ in distribution



Limit theorems

Markov and Chebyshev's inequalities

Convergence of random variables

Limit theorems

Conditional probabilities

Conditional expectation

Sum of independent identically distributed RVs

- Independent identically distributed (i.i.d.) RVs $X_1, X_2, \ldots, X_n, \ldots$
- Mean $\mathbb{E}[X_n] = \mu$ and variance $\mathbb{E}\left[(X_n - \mu)^2\right] = \sigma^2$ for all n
- Q: What happens with sum $S_N := \sum_{n=1}^{N} X_n$ as N grows?

- Expected value of sum is $\mathbb{E}[S_N] = N\mu \ \Rightarrow$ Diverges if $\mu \neq 0$
- Variance is $\mathbb{E}\left[(S_N - N\mu)^2\right] = N\sigma^2$

  $\Rightarrow$ Diverges if $\sigma \neq 0$ (always true unless X_n is a constant, boring)

- One interesting normalization $\Rightarrow \bar{X}_N := (1/N)\sum_{n=1}^{N} X_n$
- Now $\mathbb{E}\left[\bar{X}_N\right] = \mu$ and $\text{var}\left[\bar{X}_N\right] = \sigma^2/N$

  $\Rightarrow$ Law of large numbers (weak and strong)

- Another interesting normalization $\Rightarrow Z_N := \dfrac{\sum_{n=1}^{N} X_n - N\mu}{\sigma\sqrt{N}}$
- Now $\mathbb{E}[Z_N] = 0$ and $\text{var}[Z_N] = 1$ for all values of N

  $\Rightarrow$ Central limit theorem

Law of large numbers

- Sequence of i.i.d. RVs $X_1, X_2, \ldots, X_n, \ldots$ with mean μ
- Define sample average $\bar{X}_N := (1/N) \sum_{n=1}^{N} X_n$

Theorem (Weak law of large numbers)

Sample average $\bar{X}_N$ of i.i.d. sequence converges in prob. to $\mu = \mathbb{E}[X_n]$

$$\lim_{N\to\infty} \mathrm{P}\left(|\bar{X}_N - \mu| < \epsilon\right) = 1, \qquad \textit{for all } \epsilon > 0$$

Theorem (Strong law of large numbers)

Sample average $\bar{X}_N$ of i.i.d. sequence converges a.s. to $\mu = \mathbb{E}[X_n]$

$$\mathrm{P}\left(\lim_{N\to\infty} \bar{X}_N = \mu\right) = 1$$

- Strong law implies weak law. Can forget weak law if so wished

Proof of weak law of large numbers

- Weak law of large numbers is very simple to prove

Proof.

- Variance of $\bar{X}_N$ vanishes for N large

$$\text{var}\left[\bar{X}_N\right] = \frac{1}{N^2}\sum_{n=1}^{N}\text{var}\left[X_n\right] = \frac{\sigma^2}{N} \to 0$$

- But, what is the variance of $\bar{X}_N$?

$$0 \leftarrow \frac{\sigma^2}{N} = \text{var}\left[\bar{X}_N\right] = \mathbb{E}\left[(\bar{X}_N - \mu)^2\right]$$

- Then, $\bar{X}_N$ converges to μ in mean-square sense

  $\Rightarrow$ Which implies convergence in probability $\square$

- Strong law is a little more challenging. Will not prove it here

Coming full circle

- Repeated experiment $\Rightarrow$ Sequence of i.i.d. RVs $X_1, X_2, \ldots, X_n, \ldots$
  $\Rightarrow$ Consider an event of interest $X \in E$. Ex: coin comes up 'H'
- Fraction of times $X \in E$ happens in N experiments is

$$\bar{X}_N = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}\left\{X_n \in E\right\}$$

- Since the indicators also i.i.d., the strong law asserts that

$$\lim_{N \to \infty} \bar{X}_N = \mathbb{E}\left[\mathbb{I}\left\{X_1 \in E\right\}\right] = \mathrm{P}\left(X_1 \in E\right) \quad a.s.$$

- Strong law consistent with our intuitive notion of probability
  $\Rightarrow$ Relative frequency of occurrence of an event in many trials
  $\Rightarrow$ Justifies simulation-based prob. estimates (e.g. histograms)

Central limit theorem (CLT)

Theorem (Central limit theorem)

Consider a sequence of i.i.d. RVs $X_1, X_2, \ldots, X_n, \ldots$ with mean $\mathbb{E}[X_n] = \mu$ and variance $\mathbb{E}\left[(X_n - \mu)^2\right] = \sigma^2$ for all n. Then

$$\lim_{N\to\infty} \mathrm{P}\left(\frac{\sum_{n=1}^{N} X_n - N\mu}{\sigma\sqrt{N}} \leq x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du$$

- Former statement implies that for N sufficiently large

$$Z_N := \frac{\sum_{n=1}^{N} X_n - N\mu}{\sigma\sqrt{N}} \sim \mathcal{N}(0, 1)$$

$\Rightarrow$ Z_N converges in distribution to a standard normal RV

$\Rightarrow$ Remarkable universality. Distribution of X_n arbitrary

CLT (continued)

- Equivalently can say $\Rightarrow \sum_{n=1}^{N} X_n \sim \mathcal{N}(N\mu, N\sigma^2)$
- Sum of large number of i.i.d. RVs has a normal distribution

  $\Rightarrow$ Cannot take a meaningful limit here

  $\Rightarrow$ But intuitively, this is what the CLT states

Example

- Binomial RV X with parameters (n, p)
- Write as $X = \sum_{i=1}^{n} X_i$ with X_i i.i.d. Bernoulli with parameter p
- Mean $\mathbb{E}[X_i] = p$ and variance $\text{var}[X_i] = p(1-p)$

  $\Rightarrow$ For sufficiently large n $\Rightarrow X \sim \mathcal{N}(np, np(1-p))$

Conditional probabilities

Conditional pmf and cdf for discrete RVs

- Recall definition of conditional probability for events E and F

$$P(E \mid F) = \frac{P(E \cap F)}{P(F)}$$

$\Rightarrow$ Change in likelihoods when information is given, renormalization

- **Def:** Conditional pmf of RV X given Y is (both RVs discrete)

$$p_{X|Y}(x \mid y) = \mathrm{P}\left(X = x \mid Y = y\right) = \frac{\mathrm{P}\left(X = x, Y = y\right)}{\mathrm{P}\left(Y = y\right)}$$

- Which we can rewrite as

$$p_{X|Y}(x \mid y) = \frac{\mathrm{P}\left(X = x, Y = y\right)}{\mathrm{P}\left(Y = y\right)} = \frac{p_{XY}(x, y)}{p_Y(y)}$$

$\Rightarrow$ Pmf for RV X, given parameter y ("Y not random anymore")

- **Def:** Conditional cdf is (a range of X conditioned on a value of Y)

$$F_{X|Y}(x \mid y) = \mathrm{P}\left(X \leq x \mid Y = y\right) = \sum_{z \leq x} p_{X|Y}(z \mid y)$$

Conditional pmf example

- Consider independent Bernoulli RVs Y and Z, define $X = Y + Z$
- Q: Conditional pmf of X given Y? For $X = 0$, $Y = 0$

$$p_{X|Y}(X = 0 \,\big|\, Y = 0) = \frac{\mathrm{P}\left(X = 0, Y = 0\right)}{\mathrm{P}\left(Y = 0\right)} = \frac{(1-p)^2}{1-p} = 1 - p$$

- Or, from joint and marginal pmfs (just a matter of definition)

$$p_{X|Y}(X = 0 \,\big|\, Y = 0) = \frac{p_{XY}(0,0)}{p_Y(0)} = \frac{(1-p)^2}{1-p} = 1 - p$$

- Can compute the rest analogously

$$\begin{array}{lll} p_{X|Y}(0|0) = 1 - p, & p_{X|Y}(1|0) = p, & p_{X|Y}(2|0) = 0 \\ p_{X|Y}(0|1) = 0, & p_{X|Y}(1|1) = 1 - p, & p_{X|Y}(2|1) = p \end{array}$$

Conditioning on sum of Poisson RVs

- Consider independent Poisson RVs Y and Z, parameters λ_1 and λ_2
- Define $X = Y + Z$. Q: Conditional pmf of Y given X?

$$p_{Y|X}(Y = y \,\big|\, X = x) = \frac{\mathrm{P}\left(Y = y, X = x\right)}{\mathrm{P}\left(X = x\right)} = \frac{\mathrm{P}\left(Y = y\right)\mathrm{P}\left(Z = x - y\right)}{\mathrm{P}\left(X = x\right)}$$

- Used Y and Z independent. Now recall X is Poisson, $\lambda = \lambda_1 + \lambda_2$

$$\begin{aligned} p_{Y|X}(Y = y \,\big|\, X = x) &= \frac{e^{-\lambda_1}\lambda_1^y}{y!}\frac{e^{-\lambda_2}\lambda_2^{x-y}}{(x-y)!}\left[\frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^x}{x!}\right]^{-1} \\ &= \frac{x!}{y!(x-y)!}\frac{\lambda_1^y\lambda_2^{x-y}}{(\lambda_1+\lambda_2)^x} \\ &= \binom{x}{y}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^y\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{x-y} \end{aligned}$$

$\Rightarrow$ Conditioned on $X = x$, Y is binomial $(x, \lambda_1/(\lambda_1 + \lambda_2))$

Conditional pdf and cdf for continuous RVs

- **Def:** Conditional pdf of RV X given Y is (both RVs continuous)

$$f_{X|Y}(x \,|\, y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

- For motivation, define intervals $\Delta x = [x, x+dx]$ and $\Delta y = [y, y+dy]$
  $\Rightarrow$ Approximate conditional probability $\mathrm{P}\left(X \in \Delta x \,\middle|\, Y \in \Delta y\right)$ as

$$\mathrm{P}\left(X \in \Delta x \,\middle|\, Y \in \Delta y\right) = \frac{\mathrm{P}\left(X \in \Delta x, Y \in \Delta y\right)}{\mathrm{P}\left(Y \in \Delta y\right)} \approx \frac{f_{XY}(x, y)dxdy}{f_Y(y)dy}$$

- From definition of conditional pdf it follows

$$\mathrm{P}\left(X \in \Delta x \,\middle|\, Y \in \Delta y\right) \approx f_{X|Y}(x \,|\, y)dx$$

  $\Rightarrow$ What we would expect of a density

- **Def:** Conditional cdf is $\Rightarrow F_{X|Y}(x) = \int_{-\infty}^{x} f_{X|Y}(u \,|\, y)du$

Communications channel example

- Random message (RV) Y, transmit signal y (realization of Y)
- Received signal is $x = y + z$ (z realization of random noise)

  $\Rightarrow$ Model communication system as a relation between RVs

$$X = Y + Z$$

  $\Rightarrow$ Model additive noise as $Z \sim \mathcal{N}(0, \sigma^2)$ independent of Y

- Q: Conditional pdf of X given Y? Try the definition

$$f_{X|Y}(x \mid y) = \frac{f_{XY}(x, y)}{f_Y(y)} = \frac{?}{f_Y(y)}$$

  $\Rightarrow$ Problem is we don't know $f_{XY}(x, y)$. Have to calculate

- Computing conditional probs. typically easier than computing joints

Communications channel example (continued)

- If $Y = y$ is given, then "Y not random anymore"

  $\Rightarrow$ It is still random in reality, we are thinking of it as given

- If Y were not random, say $Y = y$ with y given then $X = y + Z$

  $\Rightarrow$ Cdf of X given $Y = y$ now easy (use Y and Z independent)

$$\mathrm{P}\left(X \leq x \,\middle|\, Y = y\right) = \mathrm{P}\left(y + Z \leq x | Y = y\right) = \mathrm{P}\left(Z \leq x - y\right)$$

- But since Z is normal with zero mean and variance σ^2

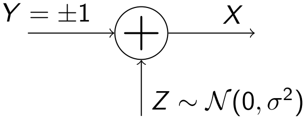
$$\begin{aligned} \mathrm{P}\left(X \leq x \,\middle|\, Y = y\right) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x-y} e^{-z^2/2\sigma^2}\, dz \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-(z-y)^2/2\sigma^2}\, dz \end{aligned}$$

$\Rightarrow$ X given $Y = y$ is normal with mean y and variance σ^2

Digital communications channel

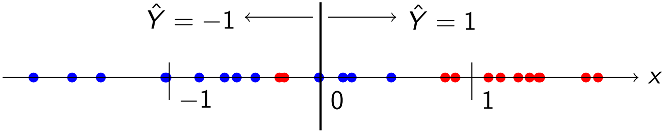
- Conditioning is a common tool to compute probabilities

- Message 1 (w.p. p) $\Rightarrow$ Transmit $Y = 1$
- Message 2 (w.p. q) $\Rightarrow$ Transmit $Y = -1$
- Received signal $\Rightarrow X = Y + Z$



- Decoding rule $\Rightarrow \hat{Y} = 1$ if $X \geq 0$, $\hat{Y} = -1$ if $X < 0$

  $\Rightarrow$ **Errors:** • to the left of 0 and • to the right



- Q: What is the probability of error, $P_e := \mathrm{P}\left(\hat{Y} \neq Y\right)$?

Output pdf

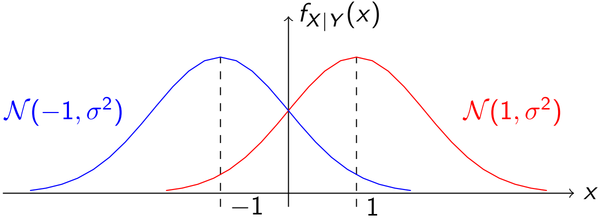
- From communications channel example we know

  $\Rightarrow$ If $Y = 1$ then $X \,\big|\, Y = 1 \sim \mathcal{N}(1, \sigma^2)$. Conditional pdf is

$$f_{X|Y}(x \,|\, 1) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-1)^2/2\sigma^2}$$

  $\Rightarrow$ If $Y = -1$ then $X \,\big|\, Y = -1 \sim \mathcal{N}(-1, \sigma^2)$. Conditional pdf is

$$f_{X|Y}(x \,|\, -1) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x+1)^2/2\sigma^2}$$



Probability of error

- Write probability of error by conditioning on $Y = \pm 1$ (total probability)

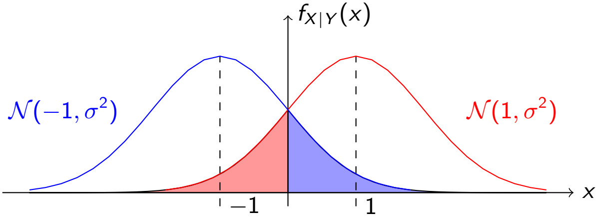
$$P_e = \mathrm{P}\left(\hat{Y} \neq Y \,\middle|\, Y = 1\right) \mathrm{P}\left(Y = 1\right) + \mathrm{P}\left(\hat{Y} \neq Y \,\middle|\, Y = -1\right) \mathrm{P}\left(Y = -1\right)$$
$$= \mathrm{P}\left(\hat{Y} = -1 \,\middle|\, Y = 1\right) \; p \qquad + \mathrm{P}\left(\hat{Y} = 1 \,\middle|\, Y = -1\right) \; q$$

- According to the decision rule

$$P_e = \mathrm{P}\left(X < 0 \,\middle|\, Y = 1\right) p + \mathrm{P}\left(X \geq 0 \,\middle|\, Y = -1\right) q$$

- But X given Y is normally distributed, then

$$P_e = \frac{p}{\sqrt{2\pi}\sigma} \int_{-\infty}^{0} e^{-(x-1)^2/2\sigma^2} dx + \frac{q}{\sqrt{2\pi}\sigma} \int_{0}^{\infty} e^{-(x+1)^2/2\sigma^2} dx$$



Conditional expectation

Markov and Chebyshev's inequalities

Convergence of random variables

Limit theorems

Conditional probabilities

Conditional expectation

Definition of conditional expectation

- **Def:** For continuous RVs X, Y, conditional expectation is

$$\mathbb{E}\left[X \,\middle|\, Y = y\right] = \int_{-\infty}^{\infty} x\, f_{X|Y}(x|y)\, dx$$

- **Def:** For discrete RVs X, Y, conditional expectation is

$$\mathbb{E}\left[X \,\middle|\, Y = y\right] = \sum_{x} x\, p_{X|Y}(x|y)$$

- Defined for given y $\Rightarrow \mathbb{E}\left[X \,\middle|\, Y = y\right]$ is a number

  $\Rightarrow$ All possible values y of Y $\Rightarrow$ random variable $\mathbb{E}\left[X \,\middle|\, Y\right]$

- $\mathbb{E}\left[X \,\middle|\, Y\right]$ a function of the RV Y, hence itself a RV

  $\Rightarrow \mathbb{E}\left[X \,\middle|\, Y = y\right]$ value associated with outcome $Y = y$

- If X and Y independent, then $\mathbb{E}\left[X \,\middle|\, Y\right] = \mathbb{E}\left[X\right]$

Conditional expectation example

- Consider independent Bernoulli RVs Y and Z, define $X = Y + Z$
- Q: What is $\mathbb{E}\left[X \,\middle|\, Y = 0\right]$? Recall we found the conditional pmf

$$\begin{aligned} &p_{X|Y}(0|0) = 1 - p, && p_{X|Y}(1|0) = p, && p_{X|Y}(2|0) = 0 \\ &p_{X|Y}(0|1) = 0, && p_{X|Y}(1|1) = 1 - p, && p_{X|Y}(2|1) = p \end{aligned}$$

- Use definition of conditional expectation for discrete RVs

$$\begin{aligned} \mathbb{E}\left[X \,\middle|\, Y = 0\right] &= \sum_x x\, p_{X|Y}(x|0) \\ &= 0 \times (1 - p) + 1 \times p + 2 \times 0 = p \end{aligned}$$

Iterated expectations

- If $\mathbb{E}\left[X \mid Y\right]$ is a RV, can compute expected value $\mathbb{E}_Y\left[\mathbb{E}_X\left[X \mid Y\right]\right]$
  Subindices clarify innermost expectation is w.r.t. X, outermost w.r.t. Y

- Q: What is $\mathbb{E}_Y\left[\mathbb{E}_X\left[X \mid Y\right]\right]$? Not surprisingly $\Rightarrow \mathbb{E}\left[X\right] = \mathbb{E}_Y\left[\mathbb{E}_X\left[X \mid Y\right]\right]$

- Show for discrete RVs (write integrals for continuous)

$$
\begin{aligned}
\mathbb{E}_Y\left[\mathbb{E}_X\left[X \mid Y\right]\right] &= \sum_y \mathbb{E}_X\left[X \mid Y = y\right] p_Y(y) = \sum_y \left[\sum_x x\, p_{X|Y}(x|y)\right] p_Y(y) \\
&= \sum_x x \left[\sum_y p_{X|Y}(x|y) p_Y(y)\right] = \sum_x x \left[\sum_y p_{XY}(x, y)\right] \\
&= \sum_x x p_X(x) = \mathbb{E}\left[X\right]
\end{aligned}
$$

- Offers a useful method to compute expected values

  $\Rightarrow$ Condition on $Y = y$ $\qquad \Rightarrow X \mid Y = y$
  $\Rightarrow$ Compute expected value over X for given y $\qquad \Rightarrow \mathbb{E}_X\left[X \mid Y = y\right]$
  $\Rightarrow$ Compute expected value over all values y of Y $\qquad \Rightarrow \mathbb{E}_Y\left[\mathbb{E}_X\left[X \mid Y\right]\right]$

Iterated expectations example

- Consider a probability class in some university
  - $\Rightarrow$ Seniors get an $A = 4$ w.p. 0.5, $B = 3$ w.p. 0.5
  - $\Rightarrow$ Juniors get a $B = 3$ w.p. 0.6, $C = 2$ w.p. 0.4
  - $\Rightarrow$ An exchange student is a senior w.p. 0.7, and a junior w.p. 0.3
- Q: Expectation of X = exchange student's grade?
- Start by conditioning on standing

$$\mathbb{E}\left[X \,\middle|\, \text{Senior}\right] = 0.5 \times 4 + 0.5 \times 3 = 3.5$$
$$\mathbb{E}\left[X \,\middle|\, \text{Junior}\right] = 0.6 \times 3 + 0.4 \times 2 = 2.6$$

- Now sum over standing's probability

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}\left[X \,\middle|\, \text{Senior}\right] \mathrm{P}(\text{Senior}) + \mathbb{E}\left[X \,\middle|\, \text{Junior}\right] \mathrm{P}(\text{Junior}) \\ &= 3.5 \times 0.7 + 2.6 \times 0.3 = 3.23\end{aligned}$$

Conditioning on sum of Poisson RVs

- Consider independent Poisson RVs Y and Z, parameters λ_1 and λ_2
- Define $X = Y + Z$. What is $\mathbb{E}\left[Y \,\middle|\, X = x\right]$?
  - $\Rightarrow$ We found $Y \,\big|\, X = x$ is binomial $(x, \lambda_1/(\lambda_1 + \lambda_2))$, hence

$$\mathbb{E}\left[Y \,\middle|\, X = x\right] = \frac{x\lambda_1}{\lambda_1 + \lambda_2}$$

- Now use iterated expectations to obtain $\mathbb{E}[Y]$
  - $\Rightarrow$ Recall X is Poisson with parameter $\lambda = \lambda_1 + \lambda_2$

$$\mathbb{E}[Y] = \sum_{x=0}^{\infty} \mathbb{E}\left[Y \,\middle|\, X = x\right] p_X(x) = \sum_{x=0}^{\infty} \frac{x\lambda_1}{\lambda_1 + \lambda_2} p_X(x)$$
$$= \frac{\lambda_1}{\lambda_1 + \lambda_2}\mathbb{E}[X] = \frac{\lambda_1}{\lambda_1 + \lambda_2}(\lambda_1 + \lambda_2) = \lambda_1$$

- Of course, since Y is Poisson with parameter λ_1

Conditioning to compute expectations

- As with probabilities conditioning is useful to compute expectations
  - $\Rightarrow$ Spreads difficulty into simpler problems (divide and conquer)

Example

- A baseball player scores X_i runs per game
  - $\Rightarrow$ Expected runs are $\mathbb{E}[X_i] = \mathbb{E}[X]$ independently of game
- Player plays N games in the season. N is random (playoffs, injuries?)
  - $\Rightarrow$ Expected value of number of games is $\mathbb{E}[N]$
- What is the expected number of runs in the season? $\Rightarrow \mathbb{E}\left[\sum_{i=1}^{N} X_i\right]$
- Both N and X_i are random, and here also assumed independent
  - $\Rightarrow$ The sum $\sum_{i=1}^{N} X_i$ is known as compound RV

Sum of random number of random quantities

Step 1: Condition on $N = n$ then

$$\left[\sum_{i=1}^{N} X_i \,\middle|\, N = n\right] = \sum_{i=1}^{n} X_i$$

Step 2: Compute expected value w.r.t. X_i, use N and the X_i independent

$$\mathbb{E}_{X_i}\left[\sum_{i=1}^{N} X_i \,\middle|\, N = n\right] = \mathbb{E}_{X_i}\left[\sum_{i=1}^{n} X_i \,\middle|\, N = n\right] = \mathbb{E}_{X_i}\left[\sum_{i=1}^{n} X_i\right] = n\mathbb{E}\left[X\right]$$

$\Rightarrow$ Third equality possible because n is a number (not a RV)

Step 3: Compute expected value w.r.t. values n of N

$$\mathbb{E}_N\left[\mathbb{E}_{X_i}\left[\sum_{i=1}^{N} X_i \,\middle|\, N\right]\right] = \mathbb{E}_N\left[N\mathbb{E}\left[X\right]\right] = \mathbb{E}\left[N\right]\mathbb{E}\left[X\right]$$

Yielding result $\Rightarrow \mathbb{E}\left[\sum_{i=1}^{N} X_i\right] = \mathbb{E}\left[N\right]\mathbb{E}\left[X\right]$

Expectation of geometric RV

Ex: Suppose X is a geometric RV with parameter p

- Calculate $\mathbb{E}[X]$ by conditioning on $Y = \mathbb{I}\{\text{"first trial is a success"}\}$
  - $\Rightarrow$ If $Y = 1$, then clearly $\mathbb{E}\left[X \mid Y = 1\right] = 1$
  - $\Rightarrow$ If $Y = 0$, independence of trials yields $\mathbb{E}\left[X \mid Y = 0\right] = 1 + \mathbb{E}[X]$
- Use iterated expectations

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}\left[X \mid Y = 1\right]\mathrm{P}(Y = 1) + \mathbb{E}\left[X \mid Y = 0\right]\mathrm{P}(Y = 0)\\ &= 1 \times p + (1 + \mathbb{E}[X]) \times (1 - p)\end{aligned}$$

- Solving for $\mathbb{E}[X]$ yields

$$\mathbb{E}[X] = \frac{1}{p}$$

- Here, direct approach is straightforward (geometric series, derivative)
  - $\Rightarrow$ Oftentimes simplifications can be major

The trapped miner example

- A miner is trapped in a mine containing three doors
- At all times $n \geq 1$ while still trapped
  - The miner chooses a door $D_n = j$, $j = 1, 2, 3$
  - Choice of door D_n made independently of prior choices
  - Equally likely to pick either door, i.e., $\mathrm{P}(D_n = j) = 1/3$
- Each door leads to a tunnel, but only one leads to safety
  - Door 1: the miner reaches safety after two hours of travel
  - Door 2: the miner returns back after three hours of travel
  - Door 3: the miner returns back after five hours of travel
- Let X denote the total time traveled till the miner reaches safety
- Q: What is $\mathbb{E}[X]$?

The trapped miner example (continued)

- Calculate $\mathbb{E}[X]$ by conditioning on first door choice D_1
  - $\Rightarrow$ If $D_1 = 1$, then 2 hours and out, i.e., $\mathbb{E}\left[X \,\middle|\, D_1 = 1\right] = 2$
  - $\Rightarrow$ If $D_1 = 2$, door choices independent so $\mathbb{E}\left[X \,\middle|\, D_1 = 2\right] = 3 + \mathbb{E}[X]$
  - $\Rightarrow$ Likewise for $D_1 = 3$, we have $\mathbb{E}\left[X \,\middle|\, D_1 = 3\right] = 5 + \mathbb{E}[X]$
- Use iterated expectations

$$\mathbb{E}[X] = \sum_{j=1}^{3} \mathbb{E}\left[X \,\middle|\, D_1 = j\right] \mathrm{P}(D_1 = j) = \frac{1}{3}\sum_{j=1}^{3} \mathbb{E}\left[X \,\middle|\, D_1 = j\right]$$

$$= \frac{2 + 3 + \mathbb{E}[X] + 5 + \mathbb{E}[X]}{3} = \frac{10 + 2\mathbb{E}[X]}{3}$$

- Solving for $\mathbb{E}[X]$ yields

$$\mathbb{E}[X] = 10$$

- You will solve it again using compound RVs in the homework

Conditional variance formula

- **Def:** The conditional variance of X given $Y = y$ is

$$\begin{aligned}\text{var}\left[X|Y=y\right] &= \mathbb{E}\left[(X - \mathbb{E}\left[X \,\middle|\, Y=y\right])^2 \,\middle|\, Y=y\right] \\ &= \mathbb{E}\left[X^2 \,\middle|\, Y=y\right] - (\mathbb{E}\left[X \,\middle|\, Y=y\right])^2\end{aligned}$$

$\Rightarrow$ $\text{var}\left[X|Y\right]$ a function of RV Y, value for $Y = y$ is $\text{var}\left[X|Y=y\right]$

- Calculate $\text{var}\left[X\right]$ by conditioning on $Y = y$. Quick guesses?

$\Rightarrow$ $\text{var}\left[X\right] \neq \mathbb{E}_Y[\text{var}_X(X \,|\, Y)]$

$\Rightarrow$ $\text{var}\left[X\right] \neq \text{var}_Y[\mathbb{E}_X(X \,|\, Y)]$

- Neither. Following conditional variance formula is the correct way

$$\text{var}\left[X\right] = \mathbb{E}_Y[\text{var}_X(X \,|\, Y)] + \text{var}_Y[\mathbb{E}_X(X \,|\, Y)]$$

Conditional variance formula (continued)

Proof.

- Start from the first summand, use linearity, iterated expectations

$$\begin{aligned}
\mathbb{E}_Y[\text{var}_X(X \mid Y)] &= \mathbb{E}_Y\left[\mathbb{E}_X(X^2 \mid Y) - (\mathbb{E}_X(X \mid Y))^2\right] \\
&= \mathbb{E}_Y\left[\mathbb{E}_X(X^2 \mid Y)\right] - \mathbb{E}_Y\left[(\mathbb{E}_X(X \mid Y))^2\right] \\
&= \mathbb{E}\left[X^2\right] - \mathbb{E}_Y\left[(\mathbb{E}_X(X \mid Y))^2\right]
\end{aligned}$$

- For the second term use variance definition, iterated expectations

$$\begin{aligned}
\text{var}_Y[\mathbb{E}_X(X \mid Y)] &= \mathbb{E}_Y\left[(\mathbb{E}_X(X \mid Y))^2\right] - \left(\mathbb{E}_Y[\mathbb{E}_X(X \mid Y)]\right)^2 \\
&= \mathbb{E}_Y\left[(\mathbb{E}_X(X \mid Y))^2\right] - (\mathbb{E}[X])^2
\end{aligned}$$

- Summing up both terms yields (blue terms cancel)

$$\mathbb{E}_Y[\text{var}_X(X \mid Y)] + \text{var}_Y[\mathbb{E}_X(X \mid Y)] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = \text{var}[X]$$

$\square$

Variance of a compound RV

- Let $X_1, X_2, \ldots$ be i.i.d. RVs with $\mathbb{E}[X_1] = \mu$ and $\text{var}[X_1] = \sigma^2$
- Let N be a nonnegative integer-valued RV independent of the X_i
- Consider the compound RV $S = \sum_{i=1}^{N} X_i$. What is $\text{var}[S]$?

- The conditional variance formula is useful here
- Earlier, we found $\mathbb{E}[S|N] = N\mu$. What about $\text{var}[S|N = n]$?

$$\text{var}\left[\sum_{i=1}^{N} X_i \Big| N = n\right] = \text{var}\left[\sum_{i=1}^{n} X_i \Big| N = n\right] = \text{var}\left[\sum_{i=1}^{n} X_i\right] = n\sigma^2$$

$\Rightarrow \text{var}[S|N] = N\sigma^2$. Used independence of N and the i.i.d. X_i

- The conditional variance formula is $\text{var}[S] = \mathbb{E}\left[N\sigma^2\right] + \text{var}[N\mu]$

Yielding result $\Rightarrow \text{var}\left[\sum_{i=1}^{N} X_i\right] = \mathbb{E}[N]\,\sigma^2 + \text{var}[N]\,\mu^2$

Glossary

- Markov's inequality
- Chebyshev's inequality
- Limit of a sequence
- Almost sure convergence
- Convergence in probability
- Mean-square convergence
- Convergence in distribution
- I.i.d. random variables
- Sample average
- Centering and scaling
- Law of large numbers
- Central limit theorem
- Conditional distribution
- Communication channel
- Probability of error
- Conditional expectation
- Iterated expectations
- Expectations by conditioning
- Compound random variable
- Conditional variance